Video Indexing and Search with Event Recounting (VISER)

BBN VISER Team

Presented by

Manasvi Tickoo

TRECVID 2014 Workshop, Orlando FL

BBN VISER at TRECVID 2014

- Participated in both MED and MER tasks
- Made submissions for all event/training/system conditions (noPRF)
- Continue to build and improve upon core system work from previous years in TRECVID
  - Multi-modal feature extraction
  - Max-margin classification and multi-stage fusion
  - Fast metadata generation and reduced memory footprint
  - Robust and fast event model training and search
- Major area of focus in 2014:
  - MG, EQG, and ES modules optimization
  - Semantic Query Generation
  - Semantic Features

Raytheon
BBN Technologies

Semantics for MED and MER

- Increasing necessity in TRECVID for semantic understanding of video
  - **MER**:
    Semantic explanation of event detection
  - **MED 000Ex and SQ**:
    Video event detection from user-defined text query only; no positive examples
- Key building blocks for both MED and MER:
  - Robust multi-modal low-level features
  - Comprehensive concepts coverage

Overview

- **Semantic Query Generation**
- **Language extraction**:
  - Speech and video text
- **Audio-visual concepts**:
  - Deep Learning
  - In-domain detectors
- **System Optimization**
- **TRECVID 14 results:**
  - MED
  - MER

Semantic Query Generation

Semantic Query Generation

- Translation of the user-defined event query (name) into the system representation
- In 010Ex and 100Ex training conditions, the semantic query is augmented/modified based on event model training

Semantic Query Generation

- Generate Semantic Query automatically from free-form description of an event:
  - Use INDRI Document Retrieval System (OTS)* for mining Gigapedia and Wikepedia articles
  - Stopwords removal and lemmatization
  - Relevant vectorization based on ranked retrieval of words using TF measure
- Key points
  - Robust hierarchical model and inference net approach for retrieval
  - Powerful query modulations (Stemmed, AND, OR, Ordered etc.)
  - Scalable and Distributed retrieval

* [Metzler and Croft ’04]

Event Query Expansion and Projection

- Each modality has its own vocabulary
  - Need to express the lemmatized event query (Q) in each vocabulary (V)
- Projection procedure:

For each word v in V, **do**

    If $v \in Q$, then $\mathrm{score}(v) = 1$ **End**

    If $v \notin Q$, then

        For each $w \in Q$, **do**

            Expand w into $W = \{w_1, \dots, w_k\}$ using Gigaword*. Then,

            For each w_k in W, **do**

                $\mathrm{score}(v) \mathrel{+}= \mathrm{sim}(v, w_k)$

            End

        End

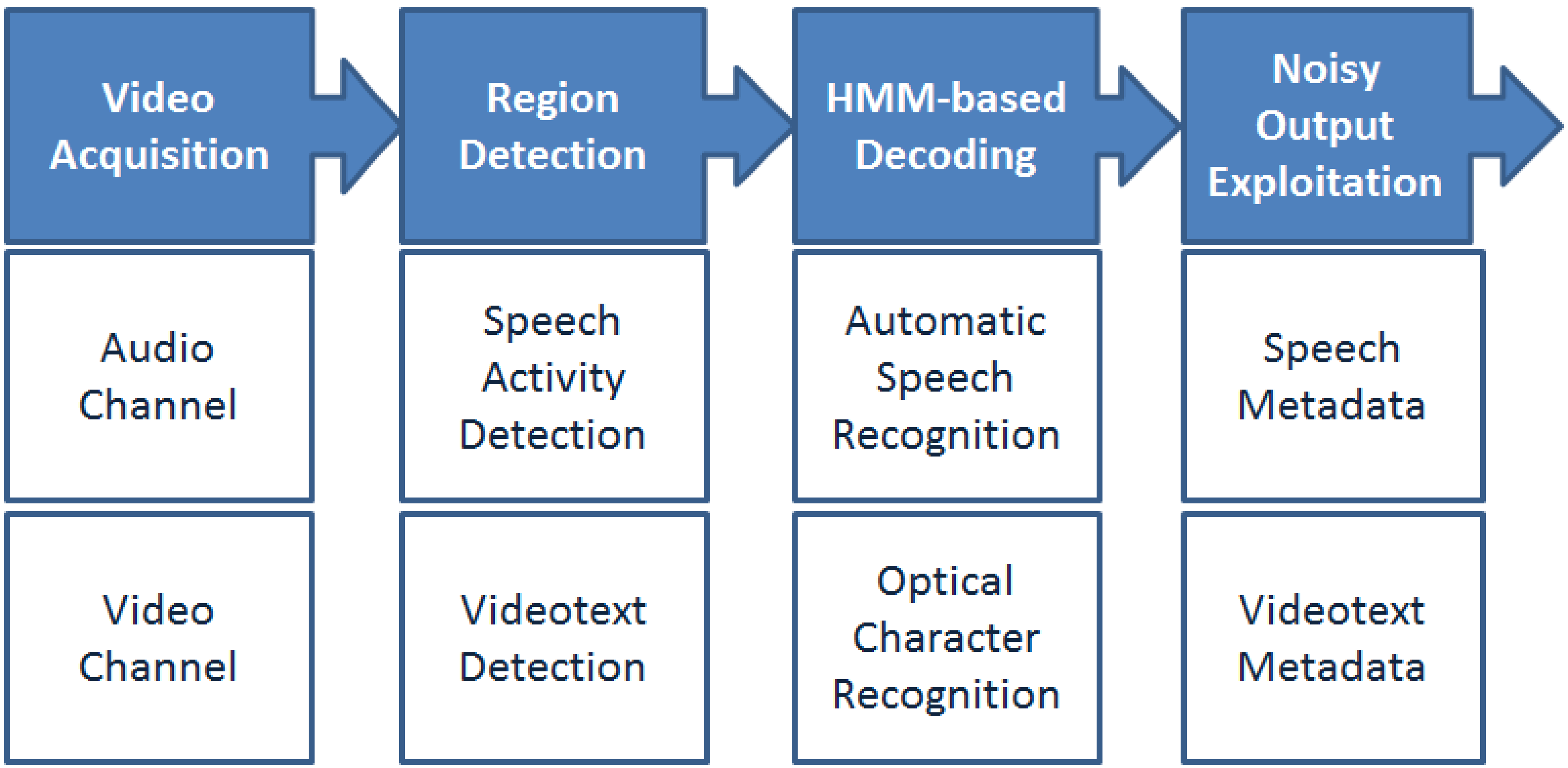
    End

End

* D. Graff, Junbo Kong, K. Chen, K. Maeda, "English Gigaword Third Edition," *Linguistic Data Consortium,* Philadelphia, 2007

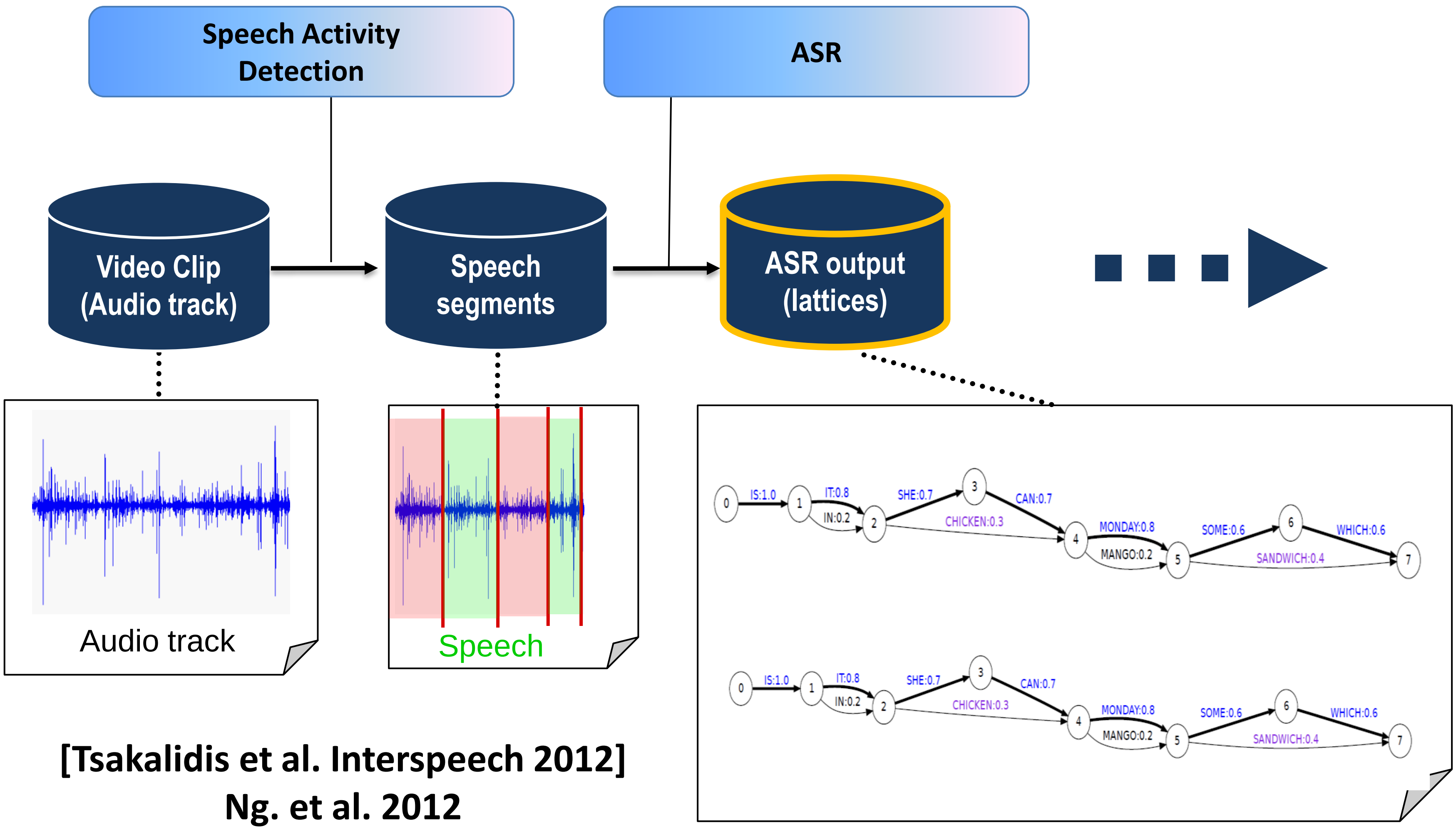
Language Extraction

Combined ASR and OCR pipelines



Wu et al. ICASSP 2014

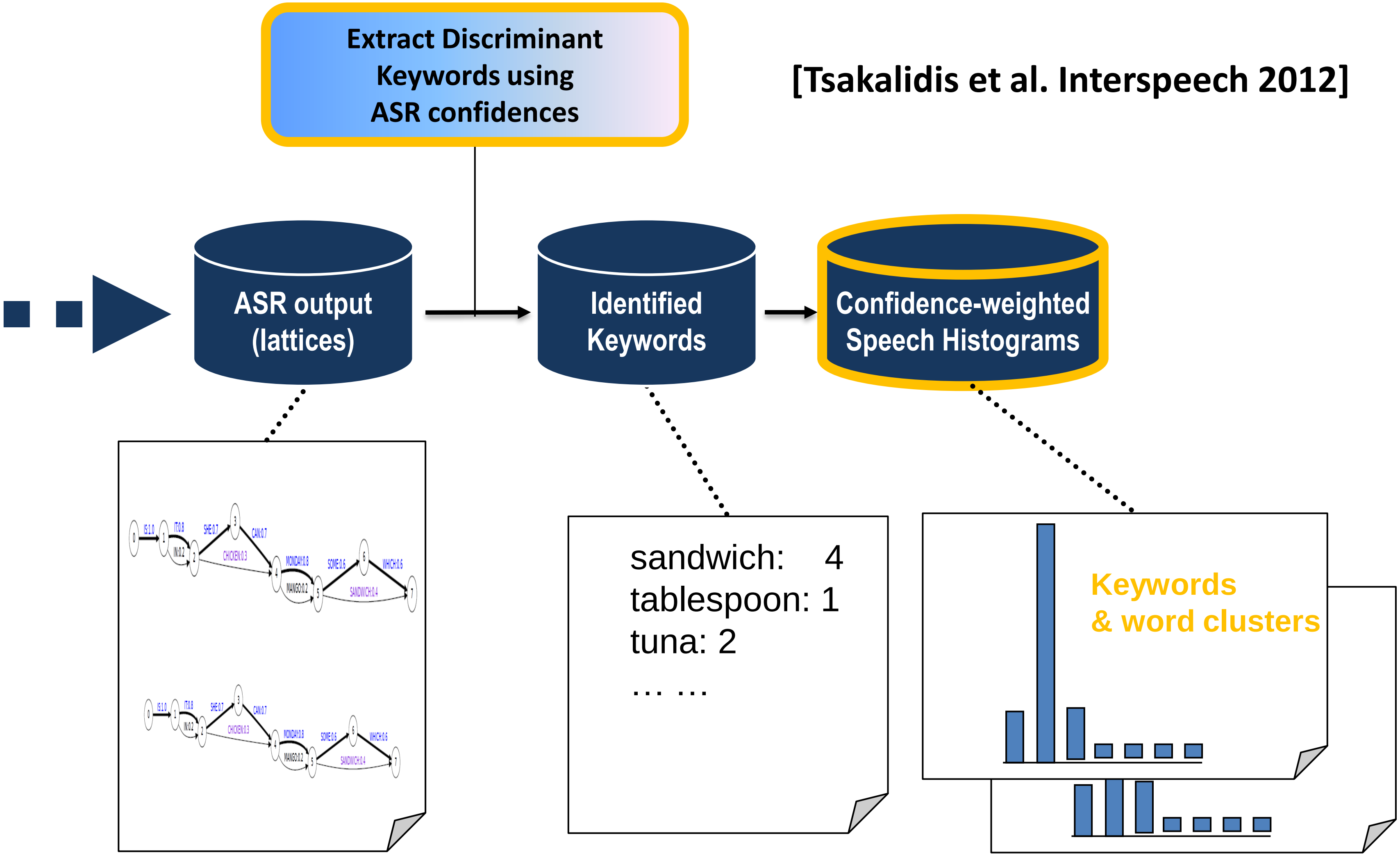
Speech



[Tsakalidis et al. Interspeech 2012]
Ng. et al. 2012

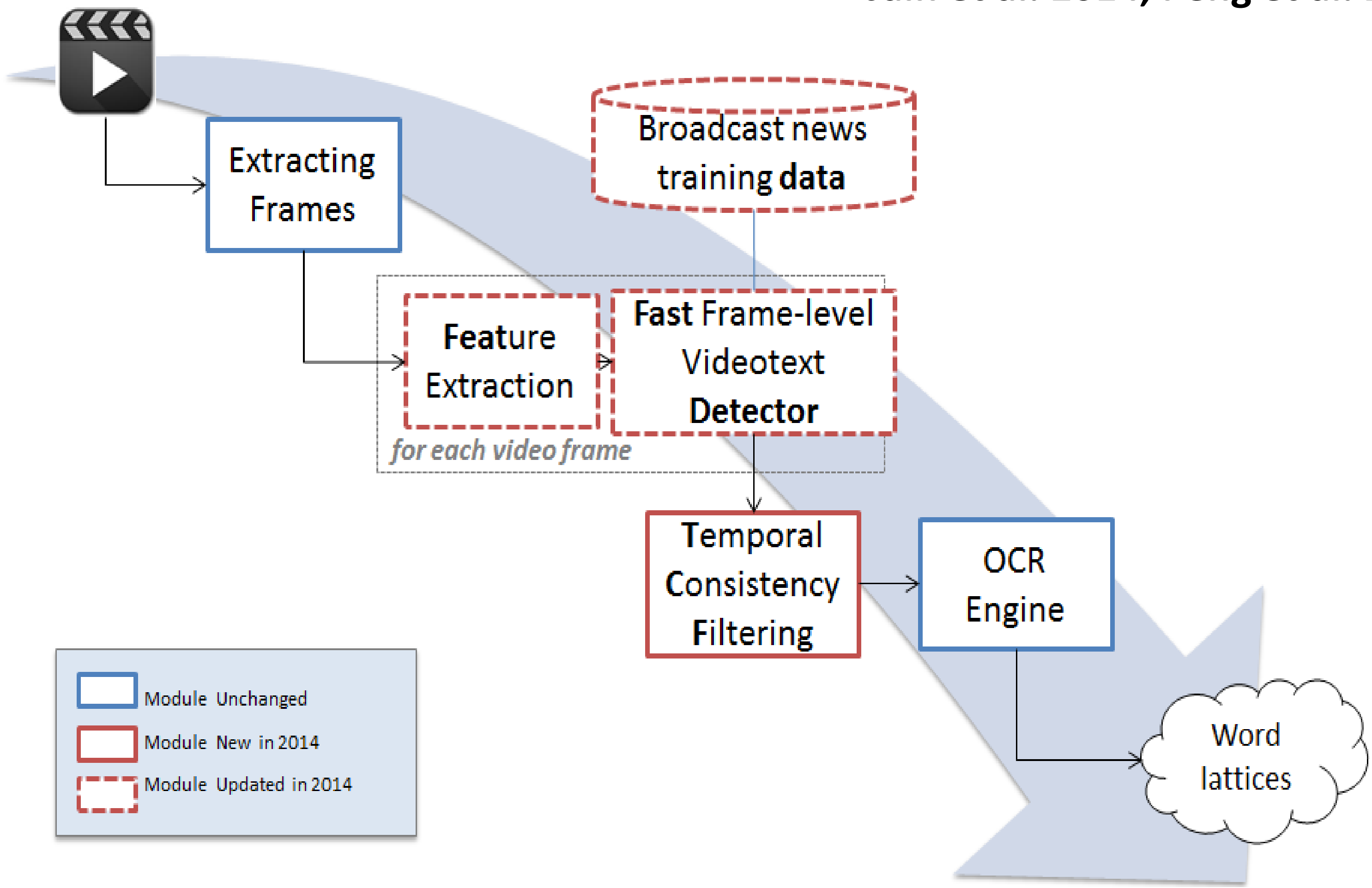
Speech (cont'd)

[Tsakalidis et al. Interspeech 2012]



Video Text

Jain et al. 2014, Peng et al. 2011



Language Content Frequency

- Keyword detections are usually precise
- Only 1/3 of the data has relevant speech, and even less has video text
- Relevant speech and text content in web video is too sparse…

Deep Learning

Deep Learning

- DCNN features trained on the ILSVRC dataset
- 8-layer DCNN on 1.2 million annotated images (GPU)
- Output layer as 1,000 dimensional semantic feature
- Last convolutional layers (fc6, fc7) as 4,096 dimensional mid-level feature for 010Ex/100Ex
- Strong performance (very close to low-level and semantic features)

Video Adaptation of ImageNet DCNN

- ImageNet DCNN output layer: 1,000 concept detectors
- Video adaptation:
  - First layer takes a 224x224x3 input image and filters it with 96 11x11x3 filters.
    - Instead of rescaling every video frames, apply the 96 filters on 10 224x224x3 rescaled sub-windows from the original video frames
  - Frame-level detection scores pooled into a single detection score for each concept
  - Spatial adaptation via spatial pyramids (SP) pooling scheme

Temporal Pooling	Spatial Pooling	Dimensions	MAP	MR0	AUC
Maximum	SP1	1,000	0.2720	0.4291	0.9345
Average	SP1	1,000	0.2735	**0.4306**	0.9344
Average	SP8	8,000	**0.2854**	0.4244	**0.9369**

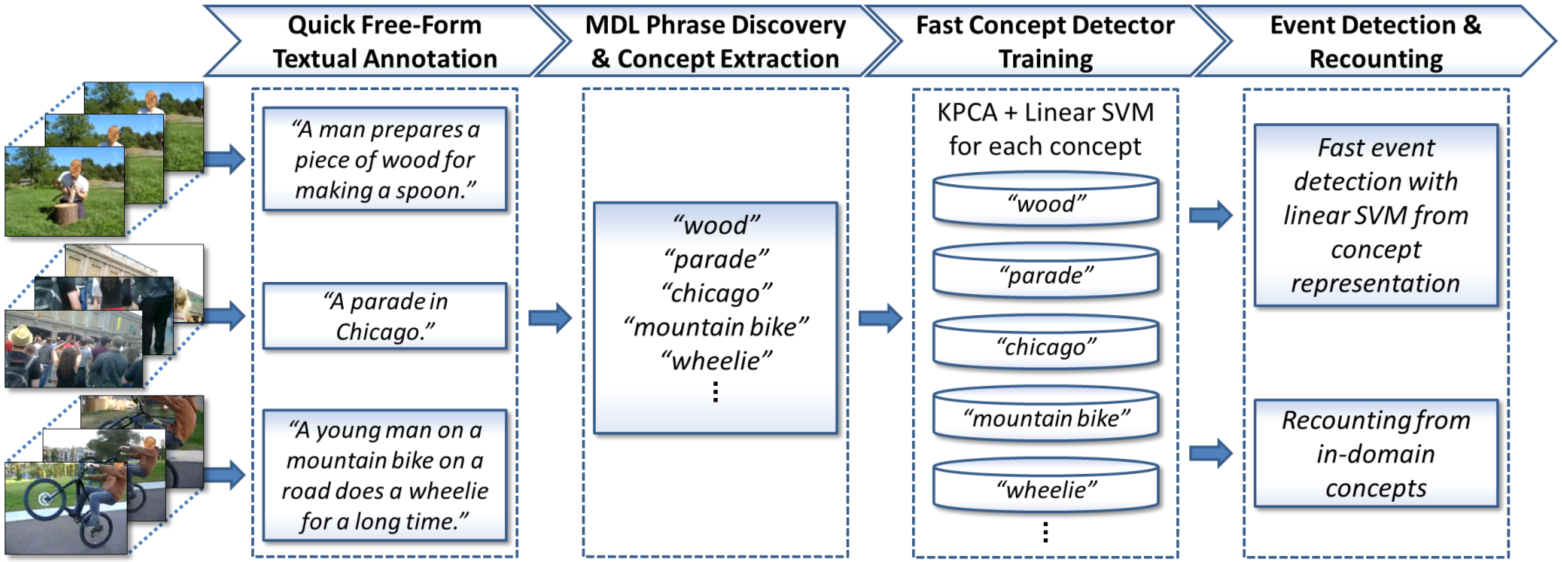
Raytheon
BBN Technologies

Audio-visual Concepts

In-domain Concept Discovery

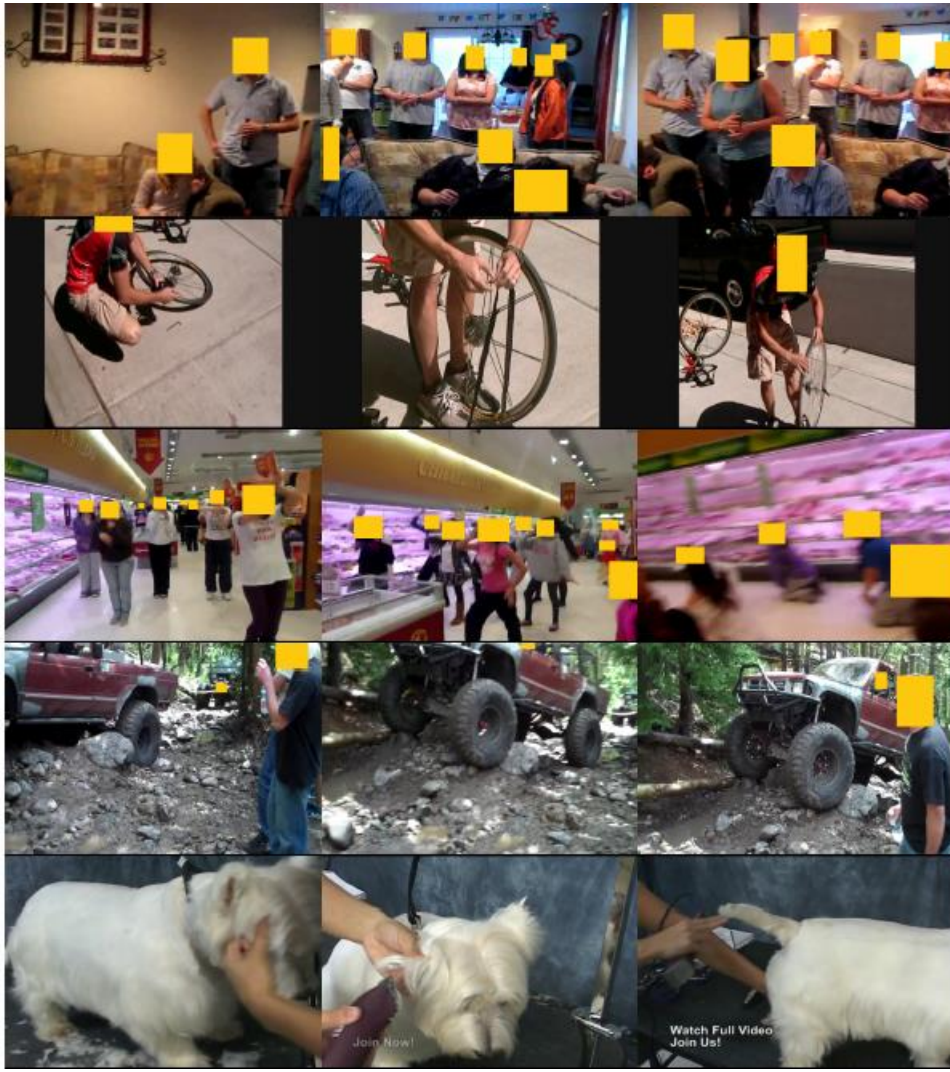
- Start with in-domain data:
  - MED research collection
- Minimized domain mismatch, but no concept annotation
- Available short text summaries in judgment files
- Discover concept labels from natural language snippets
  - Efficient to collect: **28x faster** than annotating fixed concept ontology
  - No predefined constraints on concept vocabulary (good for ad-hoc)

Weakly Supervised Concepts (WSC)



- Natural language pre-processing and phrase discovery with Minimum Description Length (MDL)
- Leverage existing MED infrastructure and extracted concept labels to train concept detectors
- Concept selection via cross-validation
- **1,800 concepts** discovered from research set and Youtube

Examples of Top Concepts Detected



E006: birthday party
WSC: piñata, people celebrate, gift
Classemes: chemical weapon, collection display setting, backpacker
Object Bank: sky, kitchen, keyboard
SUN: enclosed area, no horizon, cloth

E007: changing a vehicle tire
WSC: tire, change, replace
Classemes: chemical weapon, physical creation event, dangerous activity
Object Bank: shield, clock, basket
SUN: no horizon, manmade, enclosed area

E008: flash mob gathering
WSC: dance, flash mob, shopping
Classemes: chemical weapon, collection display setting, small group
Object Bank: keyboard, shield, kitchen
SUN: no horizon, enclosed area, cloth

E009: getting a vehicle unstuck
WSC: rocky, jeep, trail
Classemes: collection display setting, anti armor mine, mine
Object Bank: basket, shield, plate
SUN: no horizon, light/natural light, manmade

E010: grooming an animal
WSC: dog, carve, bathe
Classemes: chemical weapon, collection display setting, single doer action
Object Bank: keyboard, beach, pot
SUN: no horizon, enclosed area, manmade

WSC Concept Flexibility

- Can be trained on top of any features/modalities already present in the traditional MED infrastructure
- Can be trained with weakly annotated web data
- Can utilize visual and audio features with the same discovered labels, as well as multi-modal detectors
- Weak annotations only contain most relevant information to summarize video
  - Detectors capture **relevant** video content, not every instance of an object
- No distinction between objects/scenes/actions or word senses
  - Training process is robust enough to automatically determine best modality/most common sense

Temporal Concept Localization (Recounting)

- Video-level training, but segment-level detection
  - Apply detectors on features extracted from video segment excerpts
  - Enables rough temporal localization
    - Sliding window approach can improve temporal resolution (~10s)

Food
Eat
Chopstick
Sushi

Man
Clean
Video
Wooden
Aluminum
Wash
Laptop

Game
Boy
Play
Birthday
Kid
Playdough

Decoration
DDR
Birthday

Bake
Clean
Rink
Time-lapse

Boy
Playdough
Play
Woman
Dad

Play
Child
People
Event
Game

Woman
Presentation
Arrive

Baby
Boy
Happy
Eat
Alarm

System Optimization

Non-Linear Kernel Approximations

- Non-linear SVMs **more powerful** than linear SVMs
- Non-linear SVMs **much more expensive** at test time!
  - Linear SVM: single dot product

$$s_i = \boldsymbol{w}^T \boldsymbol{f}_i$$

  - Non-linear SVM: dot product for all margin points

On average, there are 1,200 margin points for 5,000 training videos

$$s_i = \sum_j a_j y_j K(\boldsymbol{f}_i, \boldsymbol{f}_j)$$

Linear vs. non-linear SVM for a semantic feature (100Ex)

Kernel Type	MAP	Test Time (sec/100 videos)
Linear	0.2451	**0.08**
Intersect (Non-Linear)	**0.3071**	96.0

Non-Linear Kernel Approximations

- Certain non-linear kernels can be approximated with a linear feature mapping [Vedaldi2010]
  - **Homogenous, additive** kernels: χ^2, Intersect, Hellinger's
  - Projection from $\mathbb{R}^n \to \mathbb{R}^{kn}$, where $k \leq 5$
  - Technique based on Fourier sampling theorem
- After mapping, a standard linear SVM can be used

Linear vs. non-linear vs. approximation SVM for a semantic feature (100Ex)

Kernel Type	MAP	Test Time (sec/1000 videos)
Linear	0.2451	**0.08**
Intersect (Non-Linear)	**0.3071**	96.0
Approx. Intersect (Linear)	0.3059	0.40

Feature Compression

- Up to this year:
  - All features stored in floating-point format **(4Bytes/dimension)**
- Is floating-point precision necessary?
  - Answer: **Not really**
  - Full precision feature vectors can be compressed and stored as unsigned char values **(1Byte/dimension)**

$$\boldsymbol{f}_{\mathrm{uchar}} = [a\boldsymbol{f}_{\mathrm{float}} + b], \text{ where } \begin{cases} a = \dfrac{255}{\max(\boldsymbol{f}_{\mathrm{float}}) - \min(\boldsymbol{f}_{\mathrm{float}})} \\ b = \dfrac{255 \min(\boldsymbol{f}_{\mathrm{float}})}{\min(\boldsymbol{f}_{\mathrm{float}}) - \max(\boldsymbol{f}_{\mathrm{float}})} \end{cases}$$

  - At EGQ and ES time, convert back to float and rescale
  - I/O time reduced by a factor of 4 w/o significant loss in performance
  - Total size of metadata store: **~100GB for 2T of videos!**

2013 vs. 2014: Metadata Store Comparison

Features Comparison				
	2013 System		2014 System	
Feature Type	Counts	Total Size per Video (KB)	Counts	Total Size per Video (KB)
Appearance	2	2,097	1	97
Color	1	2,097	1	65
Motion	3	6,291	1	100
Audio	3	655	1	46
Deep Learning	0	N/A	2	26
Semantic	6	6	9	24
Language	2	176	2	28
SUM	**17**	**11,322**	**17**	**386**

- 2013 system:
  - Metadata generation takes over a **month for 100,000 videos** on a cluster of computers
- 2014 system:
  - Metadata generation takes around **10 days for 200,000 videos** on the same cluster of computer

Raytheon
BBN Technologies

TRECVID 14 Results

MED Performance

Pre-specified

	MAP	MR0
100Ex	29.8%	56.3%
010Ex	18.0%	41.7%
000Ex	5.7%	24.3%
SQ	5.3%	20.3%

Ad Hoc

	MAP	MR0
100Ex	22.6%	46.9%
010Ex	10.9%	33.3%
000Ex	3.7%	14.7%
SQ	3.1%	11.7%

- Consistent pre-specified and ad hoc performance
  - Our in-domain and deep learning concepts are event-independent and generalize well to different event queries
- Strong overall performance in all system conditions

Running Times

Event Query (Median Processing Time) Single COTS machine	
SQ	3.5 min
000Ex	1.4 min
010Ex	7.7 min
100Ex	28.3 min

Event Search (Median Processing Time) Single COTS machine	
SQ	1.9 min
000Ex	1.9 min
010Ex	1.8 min
100Ex	1.5 min

- One of the fastest systems for SQ, EQG, ES while maintaining strong performance
- Metadata generation takes only 0.027 hours per hour of video (i.e. 1/35 of the playback time)

MER Approach

- Detect concept instances from various modalities
- Aggregate detections by modality, based on the initial event-specific semantic query
- Generate a human-readable recounting containing itemized detections along with confidence and relevance information

MER Results

- 5 human judges
- Query Conciseness:
  - 17 % strongly agree (highest)
  - 59 % agree votes
- Key evidence convincing:
  - Lowest strongly disagree (7%)
  - Highest strongly agree (27%)

Summary

- Reliable semantic extraction from video is key for all MED/MER tasks
- Multi-modal combination of semantic information is especially important
- Semantics can now match low-level feature performance in 010Ex/100Ex MED
- Careful feature design leads to much smaller metadata store, and thus faster MG, EQG and ES
- Nonlinear kernel approximation achieves good performance at reduced computational cost

Acknowledgement

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

Thank You!